

Approximation Algorithms for Bregman Co-clustering and Tensor Clustering*

Stefanie Jegelka
MPI for Biological Cybernetics
72070 Tübingen, Germany

Suvrit Sra
MPI for Biological Cybernetics
72070 Tübingen, Germany

Arindam Banerjee
Univ. of Minnesota
MN 55455, USA

Abstract

In the past few years powerful generalizations to the Euclidean k-means problem have been made, such as Bregman clustering [7], co-clustering (i.e., simultaneous clustering of rows and columns of an input matrix) [9, 18], and tensor clustering [8, 34]. Like k-means, these more general problems also suffer from the NP-hardness of the associated optimization. Researchers have developed approximation algorithms of varying degrees of sophistication for k-means, k-medians, and more recently also for Bregman clustering [2]. However, there seem to be no approximation algorithms for Bregman co- and tensor clustering. In this paper we derive the first (to our knowledge) guaranteed methods for these increasingly important clustering settings. Going beyond Bregman divergences, we also prove an approximation factor for tensor clustering with arbitrary separable metrics. Through extensive experiments we evaluate the characteristics of our method, and show that it also has practical impact.

1 Introduction

Partitioning data points into clusters is a fundamentally hard problem. The well-known Euclidean k-means problem that partitions the input data points (vectors in \mathbb{R}^d) into K clusters while minimizing sums of their squared distances to corresponding cluster centroids, is an NP hard problem [19] (exponential in d). However, simple and frequently used procedures that rapidly obtain local minima exist since a long time [23, 28].

Because of its wide applicability and importance, the Euclidean k-means problem has been generalized in several directions. Specific examples relevant to this paper include:

- *Bregman clustering* [7], where instead of minimizing squared Euclidean distances one minimizes Bregman divergences (which are generalized distance functions, see (3.10) or [13] for details),
- *Bregman co-clustering* [9] (which includes both Euclidean [16] and information-theoretic co-clustering [18] as special cases), where the set of input vectors is viewed as a matrix and one *simultaneously* clusters rows and columns to obtain coherent submatrices (co-clusters), while minimizing a Bregman divergence, and
- *Tensor clustering* or multiway clustering [34], especially the version based on Bregman divergences [8], where one simultaneously clusters along various dimensions of the input tensor.

For these problems too, the commonly used heuristics perform well, but do not provide theoretical guarantees (or at best assure local optimality). For k-means type clustering problems—i.e., problems

*A part of the theory of this paper appeared in MPI-TR-#177 [35].

that group together input vectors into clusters while minimizing “distance” to cluster centroids—there exist several algorithms that approximate a globally optimal solution. We refer the reader to [1, 2, 6, 27], and the numerous references therein for more details.

In stark contrast, approximation algorithms for tensor clustering are much less studied. We are aware of only two very recent attempts (both papers are from 2008) for the two-dimensional special case of co-clustering, namely, [4] and [31]—and both of the papers follow similar approaches to obtain their approximation guarantees. Both prove a $2\alpha_1$ -approximation for Euclidean co-clustering, Puolamäki et al. [31] an additional factor of $(1 + \sqrt{2})$ for binary matrices and an ℓ_1 norm objective, and Anagnostopoulos et al. [4] a factor of $3\alpha_1$ for co-clustering real matrices with ℓ_p norms. In all factors α_1 is an approximation guarantee for clustering either rows or columns. In this paper, we build upon [4] and obtain approximation algorithms for tensor clustering with Bregman divergences and arbitrary separable metrics such as ℓ_p -norms. The latter result is of particular interest for ℓ_1 -norm based tensor clustering, which may be viewed as a generalization of k-medians to tensors. In the terminology of [7], we focus on the “block average” versions of co- and tensor clustering.

Additional discussion and relevant references for co-clustering can be found in [9], while for the lesser known problem of tensor clustering more background can be gained by referring to [3, 8, 10, 21, 29, 34].

1.1 Contributions

The main contribution of this paper is the analysis of an approximation algorithm for tensor clustering that achieves an approximation ratio of $O(m\alpha)$, where m is the order of the tensor and α is the approximation factor of a corresponding 1D clustering algorithm. Our results apply to a fairly broad class of objective functions, including metrics such as ℓ_p norms or Hilbertian metrics [24, 33], and divergence functions such as Bregman divergences [13] (with some assumptions). As corollaries, our results solve two open problems posed by [4], viz., whether their methods for Euclidean co-clustering could be extended to Bregman co-clustering, and if one could extend the approximation guarantees to tensor clustering. Owing to the structure of the algorithm, our results also give insight into properties of the tensor clustering problem as such, namely, a bound on the amount of information inherent in the joint consideration of several dimensions.

In addition, we provide extensive experimental validation of the theoretical claims, which forms an additional contribution of this paper.

2 Background

Traditionally, “center” based clustering algorithms seek partitions of columns of an input matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ into clusters $\mathcal{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_K\}$, and find “centers” $\boldsymbol{\mu}_k$ that minimize the objective

$$J(\mathcal{C}) = \sum_{k=1}^K \sum_{\mathbf{x} \in \mathcal{C}_k} d(\mathbf{x}, \boldsymbol{\mu}_k), \quad (2.1)$$

where the function $d(\mathbf{x}, \mathbf{y})$ measures cluster quality. The “center” $\boldsymbol{\mu}_k$ of cluster \mathcal{C}_k is given by the mean of the points in \mathcal{C}_k when $d(\mathbf{x}, \mathbf{y})$ is a Bregman divergence [7]. Co-clustering extends (2.1) to seek simultaneous partitions (and centers $\boldsymbol{\mu}_{IJ}$) of rows and columns of \mathbf{X} , so that the objective function

$$J(\mathcal{C}) = \sum_{I,J} \sum_{i \in I, j \in J} d(x_{ij}, \boldsymbol{\mu}_{IJ}), \quad (2.2)$$

is minimized; $\boldsymbol{\mu}_{IJ}$ denotes the (scalar) “center” of the cluster described by the row and column index sets, viz., I and J . Formulation (2.2) is easily generalized to tensors, as shown in Section 2.2 below. However, we first recall basic notation about tensors before formally presenting the tensor clustering problem. Tensors are well-studied in multilinear algebra [22], and they are gaining importance in both data mining and machine learning applications [26, 36].

2.1 Tensors

A large part of the material in this section is derived from the well-written paper of de Silva and Lim [17]—their notation turns out to be particularly suitable for our analysis. An order- m tensor A may be viewed as an element of the vector space $\mathbb{R}^{n_1 \times \dots \times n_m}$ (in this paper we denote matrices and tensors using sans-serif letters). An individual component of the tensor A is represented by the multiply-indexed value $a_{i_1 i_2 \dots i_m}$, where $i_j \in \{1, \dots, n_j\}$ for $1 \leq j \leq m$.

Multilinear matrix multiplication

For us the most important operation on tensors is that of multilinear matrix multiplication, which is a generalization of the familiar concept of matrix multiplication. Matrices *act* on other matrices by either left or right multiplication. For an order-3 tensor, there are three dimensions on which a matrix may act via matrix multiplication. For example, for an order-3 tensor $A \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, and three matrices $P \in \mathbb{R}^{p_1 \times n_1}$, $Q \in \mathbb{R}^{p_2 \times n_2}$, and $R \in \mathbb{R}^{p_3 \times n_3}$, *multilinear matrix multiplication* is the operation defined by the action of these three matrices on the different dimensions of A that yields the tensor $A' \in \mathbb{R}^{p_1 \times p_2 \times p_3}$. Formally, the entries of the tensor A' are given by

$$a'_{lmn} = \sum_{i,j,k=1}^{n_1, n_2, n_3} p_{li} q_{mj} r_{nk} a_{ijk}, \quad (2.3)$$

and this operation is written compactly as

$$A' = (P, Q, R) \cdot A. \quad (2.4)$$

Multilinear multiplication extends naturally to tensors of arbitrary order. If $A \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_m}$, and $P_1 \in \mathbb{R}^{p_1 \times n_1}, \dots, P_m \in \mathbb{R}^{p_m \times n_m}$, then $A' = ((P_1, \dots, P_m) \cdot A) \in \mathbb{R}^{p_1 \times \dots \times p_m}$ has components

$$a'_{i_1 i_2 \dots i_m} = \sum_{j_1, \dots, j_m=1}^{n_1, \dots, n_m} p_{i_1 j_1}^{(1)} \cdots p_{i_m j_m}^{(m)} a_{j_1 \dots j_m}, \quad (2.5)$$

where $p_{ij}^{(k)}$ denotes the ij -th entry of matrix P_k .

Example 2.1 (Matrix Multiplication). Let $A \in \mathbb{R}^{n_1 \times n_2}$, $P \in \mathbb{R}^{p \times n_1}$, and $Q \in \mathbb{R}^{q \times n_2}$ be three matrices. The matrix product PAQ^\top can be written as the multilinear multiplication $(P, Q) \cdot A$.

Proposition 2.2 (Basic Properties). *The following properties of multilinear multiplication are easily verified (and generalized to tensors of arbitrary order):*

1. **Linearity:** Let $\alpha, \beta \in \mathbb{R}$, and A and B be tensors with same dimensions, then

$$(P, Q) \cdot (\alpha A + \beta B) = \alpha (P, Q) \cdot A + \beta (P, Q) \cdot B$$

2. **Product rule:** For matrices P_1, P_2, Q_1, Q_2 of appropriate dimensions, and a tensor A

$$(P_1, P_2) \cdot ((Q_1, Q_2) \cdot A) = (P_1 Q_1, P_2 Q_2) \cdot A$$

3. **Multilinearity:** Let $\alpha, \beta \in \mathbb{R}$, and P, Q , and R be matrices of appropriate dimensions. Then, for a tensor A the following holds

$$(P, \alpha Q + \beta R) \cdot A = \alpha (P, Q) \cdot A + \beta (P, R) \cdot A$$

Vector Norms

The standard vector ℓ_p -norms can be easily extended to tensors, and are defined as

$$\|A\|_p = \left(\sum_{i_1, \dots, i_m} |a_{i_1 \dots i_m}|^p \right)^{1/p}, \quad (2.6)$$

for $p \geq 1$. In particular for $p = 2$ we get the “Frobenius” norm, also written as $\|A\|_F$.

Inner Product

The Frobenius norm induces an inner-product that can be defined as

$$\langle A, B \rangle = \sum_{i_1, \dots, i_m} a_{i_1 \dots i_m} b_{i_1 \dots i_m}, \quad (2.7)$$

so that $\|A\|_F^2 = \langle A, A \rangle$ holds as usual.

Proposition 2.3. *The following property of this inner product is easily verified (a generalization of the familiar property $\langle Ax, By \rangle = \langle x, A^\top By \rangle$ for vectors):*

$$\langle (P_1, \dots, P_m) \cdot A, (Q_1, \dots, Q_m) \cdot B \rangle = \langle A, (P_1^\top Q_1, \dots, P_m^\top Q_m) \cdot B \rangle. \quad (2.8)$$

Proof: Using definition (2.5) and the inner-product rule (2.7) we have

$$\begin{aligned} \langle (P_1, \dots, P_m) \cdot A, (Q_1, \dots, Q_m) \cdot B \rangle &= \sum_{i_1, \dots, i_m} \sum_{j_1, \dots, j_m} \sum_{k_1, \dots, k_m} p_{i_1 j_1}^{(1)} q_{i_1 k_1}^{(1)} \cdots p_{i_m j_m}^{(m)} q_{i_m k_m}^{(m)} a_{j_1 \dots j_m} b_{k_1 \dots k_m}, \\ &= \sum_{j_1, \dots, j_m} \left(\sum_{i_1} p_{i_1 j_1}^{(1)} q_{i_1 k_1}^{(1)} \right) \cdots \left(\sum_{i_m} p_{i_m j_m}^{(m)} q_{i_m k_m}^{(m)} \right) a_{j_1 \dots j_m} b_{k_1 \dots k_m} \\ &= \sum_{j_1, \dots, j_m} (P_1^\top Q_1)_{j_1 k_1} \cdots (P_m^\top Q_m)_{j_m k_m} a_{j_1 \dots j_m} b_{k_1 \dots k_m} = \sum_{j_1 \dots j_m} a_{j_1 \dots j_m} b'_{j_1 \dots j_m} = \langle A, B' \rangle, \end{aligned}$$

where $B' = (P_1^\top Q_1, \dots, P_m^\top Q_m) \cdot B$.

Divergence

Finally, we define an arbitrary *divergence* function $d(X, Y)$ between two m -dimensional tensors X, Y as an elementwise sum of individual divergences, i.e.,

$$d(X, Y) = \sum_{i_1, \dots, i_m} d(x_{i_1, \dots, i_m}, y_{i_1, \dots, i_m}), \quad (2.9)$$

and we will define the scalar divergence $d(x, y)$ as the need arises.

2.2 Tensor clustering

Let $A \in \mathbb{R}^{n_1 \times \dots \times n_m}$ be an order- m tensor that we wish to partition into coherent sub-tensors (or clusters). A basic approach is to minimize the sum of the divergences between individual (scalar) elements in each cluster to their corresponding (scalar) cluster “centers”. Readers familiar with [9] will recognize this to be a “block-average” variant of tensor clustering.

Assume that each dimension j ($1 \leq j \leq m$) is partitioned into k_j clusters. Let $C_j \in \{0, 1\}^{n_j \times k_j}$ be the cluster indicator matrix for dimension j , where the ik -th entry of such a matrix is one if and only if index i belongs to the k -th cluster ($1 \leq k \leq k_j$) for dimension j . Then, the *tensor clustering* problem is (cf. 2.2):

$$\underset{C_1, \dots, C_m, M}{\text{minimize}} \quad d(A, (C_1, \dots, C_m) \cdot M), \quad \text{s.t. } C_j \in \{0, 1\}^{n_j \times k_j}, \quad (2.10)$$

where the tensor M collects all the cluster “centers.”

3 Algorithm and Analysis

Given formulation (2.10), our algorithm, which we name **Combination Tensor Clustering** (CoTeC), follows the simple outline:

1. Cluster along each dimension j , using an approximation algorithm to obtain clustering C_j ; Let $C = (C_1, \dots, C_m)$
2. Compute $M = \operatorname{argmin}_{X \in \mathbb{R}^{k_1 \times \dots \times k_m}} d(A, C \cdot X)$.
3. Return the tensor clustering (C_1, \dots, C_m) (with representatives M).

Instead of clustering one dimension at a time, we can also cluster along t dimensions at a time, which we will call *t-dimensional clustering*. For an order- m tensor, with $t = 1$ we form groups of order- $(m - 1)$ tensors. For illustration, consider an order-3 tensor \mathbf{A} for which we group matrices when $t = 1$. For the first dimension we cluster the objects $\mathbf{A}(i, :, :)$ (using MATLAB notation) to obtain cluster indicators \mathbf{C}_1 ; we repeat the procedure for the second and third dimensions. The approximate tensor clustering will be the combination $(\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3)$. As we assumed the Bregman divergences to be separable, the sub-tensors, e.g., $\mathbf{A}(i, :, :)$ can be simply treated as vectors.

Apart from [4, 31], all approximation guarantees refer to one-dimensional clustering algorithms. Any one-dimensional approximation algorithm can be used as a base method for our scheme outlined above. For example, the method of Ackermann and Blömer [1], or the more practical Bregman clustering approaches of [30, 35]¹ are two potential choices, though with different approximation factors. Clustering along individual dimensions and then combining the results to obtain a tensor clustering might seem counterintuitive to the idea of “co”-clustering, where one *simultaneously* clusters along different dimensions. However, our analysis will show that dimension-wise clustering suffices to obtain strong approximation guarantees for tensor clustering—a fact often observed empirically too. At the same time, our Theorem 1 bounds the amount of information that can lie in the simultaneous consideration of multiple dimensions.

3.1 Results

The main contribution of this paper is the following approximation guarantee for CoTeC, which we prove in the remainder of this section.

Theorem 3.1 (Approximation). *Let \mathbf{A} be an order- m tensor and let \mathcal{C}_j denote its clustering along the j th subset of t dimensions ($1 \leq j \leq m/t$), as obtained from a multiway clustering algorithm with guarantee α_t^2 . Let $\mathcal{C} = (\mathcal{C}_1, \dots, \mathcal{C}_{m/t})$ denote the induced tensor clustering, and $J_{OPT}(m)$ the best m -dimensional clustering. Then,*

$$J(\mathcal{C}) \leq p(m/t) \rho_d \alpha_t J_{OPT}(m), \quad \text{with} \quad (3.1)$$

1. $\rho_d = 1$ and $p(m/t) = 2^{\log_2 m/t}$ if $d(x, y) = (x - y)^2$,
2. $\rho_d = 1$ and $p(m/t) = 2m/t$ if $d(x, y)$ is a metric³.

Theorem 3.1 is quite general, and it can be combined with some natural assumptions (see Section 3.5) to yield results for tensor clustering with more general divergence functions too (here ρ_d might be greater than 1).

3.2 Analysis: Theorem 3.1, Euclidean case

We begin our proof with the Euclidean case, i.e., $d(x, y) = (x - y)^2$. Our proof is inspired by the techniques of [4]. We establish that given a clustering algorithm which clusters along t of the m dimensions at a time⁴ with an approximation factor of α_t , our CoTeC algorithm achieves an objective within a factor $O(\lceil m/t \rceil \alpha_t)$ of the optimal. For example, for $t = 1$ we can use the seeding methods of [30, 35] or the stronger approximation algorithms of [1]. We assume without loss of generality (wlog) that $m = 2^h t$ for an integer h (otherwise, pad in empty dimensions).

Since for the squared Frobenius norm, each cluster “center” is given by the mean, we can recast Problem (2.10) into a more convenient form. To that end, note that the individual entries of the

¹Both [30, 35] discovered essentially the same method for Bregman clustering, though the analysis of [30] is somewhat sharper.

²We say an approximation algorithm has guarantee α if it yields a solution that achieves an objective value within a factor $O(\alpha)$ of the optimum.

³The results can be trivially extended to λ -relaxed metrics that satisfy $d(x, y) \leq \lambda(d(x, z) + d(z, y))$; the corresponding approximation factor just gets scaled by λ .

⁴One could also consider clustering differently sized subsets of the dimensions, say $\{t_1, \dots, t_r\}$, where $t_1 + \dots + t_r = m$. However, this requires unilluminating notational jugglery, which we can skip for simplicity of exposition.

means tensor \mathbf{M} are given by (cf. (2.2))

$$\mu_{I_1 \dots I_m} = \frac{1}{|I_1| \dots |I_m|} \sum_{i_1 \in I_1, \dots, i_m \in I_m} a_{i_1 \dots i_m}, \quad (3.2)$$

with index sets I_j for $1 \leq j \leq m$. Let $\bar{\mathbf{C}}_j$ be the normalized cluster indicator matrix obtained by normalizing the columns of \mathbf{C}_j , so that $\bar{\mathbf{C}}_j^\top \bar{\mathbf{C}}_j = \mathbf{I}_{k_j}$. Then, we can rewrite (2.10) in terms of projection matrices \mathbf{P}_j as:

$$\underset{\mathcal{C}=(\bar{\mathbf{C}}_1, \dots, \bar{\mathbf{C}}_m)}{\text{minimize}} \quad J(\mathcal{C}) = \|\mathbf{A} - (\mathbf{P}_1, \dots, \mathbf{P}_m) \cdot \mathbf{A}\|_{\mathbb{F}}^2, \quad \text{s.t. } \mathbf{P}_j = \bar{\mathbf{C}}_j \bar{\mathbf{C}}_j^\top. \quad (3.3)$$

Lemma 3.2 (Pythagorean). *Let $\mathbf{P} = (\mathbf{P}_1, \dots, \mathbf{P}_t)$, $\mathbf{S} = (\mathbf{P}_{t+1}, \dots, \mathbf{P}_m)$, and $\mathbf{P}^\perp = (\mathbf{I} - \mathbf{P}_1, \dots, \mathbf{I} - \mathbf{P}_t)$ be collections of projection matrices \mathbf{P}_j . Then,*

$$\|(\mathbf{P}, \mathbf{S}) \cdot \mathbf{A} + (\mathbf{P}^\perp, \mathbf{R}) \cdot \mathbf{B}\|^2 = \|(\mathbf{P}, \mathbf{S}) \cdot \mathbf{A}\|^2 + \|(\mathbf{P}^\perp, \mathbf{R}) \cdot \mathbf{B}\|^2,$$

where \mathbf{R} is a collection of $m - t$ projection matrices.

Proof. Using $\|\mathbf{A}\|_{\mathbb{F}}^2 = \langle \mathbf{A}, \mathbf{A} \rangle$ we can rewrite the l.h.s. as

$$\begin{aligned} & \|(\mathbf{P}, \mathbf{S}) \cdot \mathbf{A} + (\mathbf{P}^\perp, \mathbf{R}) \cdot \mathbf{B}\|^2 \\ &= \|(\mathbf{P}, \mathbf{S}) \cdot \mathbf{A}\|^2 + \|(\mathbf{P}^\perp, \mathbf{R}) \cdot \mathbf{B}\|^2 + 2\langle (\mathbf{P}, \mathbf{S}) \cdot \mathbf{A}, (\mathbf{P}^\perp, \mathbf{R}) \cdot \mathbf{B} \rangle. \end{aligned}$$

The last term is immediately seen to be zero using Property (2.8) and the fact that $\mathbf{P}_j^\top \mathbf{P}_j^\perp = \mathbf{P}_j(\mathbf{I} - \mathbf{P}_j) = \mathbf{0}$. \square

Some more notation: Since we cluster along t dimensions at a time, we recursively partition the initial set of all m dimensions until (after $\log(m/t) + 1$ steps), the sets of dimensions have length t . Let l denote the level of recursion, starting at $l = \log(m/t) = h$ and going down to $l = 0$. At level l , the sets of dimensions will have length $2^l t$ (so that for $l = 0$ we have t dimensions). We represent each clustering along a subset of $2^l t$ dimensions by its corresponding $2^l t$ projection matrices. We gather these projection matrices into the collection \mathbf{P}_i^l (note boldface), where the index i ranges from 1 to 2^{h-l} .

Example 3.3. Consider an order-8 tensor where we group $t = 2$ dimensions at a time. Then, $h = \log(m/t) = 2$ and we have 3 levels. We recursively divide the set of dimensions in the middle, i.e., $\{1, \dots, 8\}$ into $\{1, \dots, 4\}$ and $\{5, \dots, 8\}$ and so on, ending with $\{\{1, 2\}, \{3, 4\}, \{5, 6\}, \{7, 8\}\}$. The projection matrix for dimension i is \mathbf{P}_i , and the full tensor clustering is represented by $(\mathbf{P}_1, \dots, \mathbf{P}_8)$. For each level $l = 0, 1, 2$, individual collections of projection matrices \mathbf{P}_i^l are

$$\begin{aligned} \mathbf{P}_1^2 &= (\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4, \mathbf{P}_5, \mathbf{P}_6, \mathbf{P}_7, \mathbf{P}_8) \\ \mathbf{P}_1^1 &= (\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4), \quad \mathbf{P}_2^1 = (\mathbf{P}_5, \mathbf{P}_6, \mathbf{P}_7, \mathbf{P}_8) \\ \mathbf{P}_1^0 &= (\mathbf{P}_1, \mathbf{P}_2), \quad \dots, \quad \mathbf{P}_4^0 = (\mathbf{P}_7, \mathbf{P}_8). \end{aligned}$$

We also need some notation to represent a complete tensor clustering along all m dimensions, where *only a subset* of $2^l t$ dimensions are clustered. We pad the collection \mathbf{P}_i^l with $m - 2^l t$ identity matrices for the non-clustered dimensions, and call this padded collection \mathbf{Q}_i^l . With recursive partitioning of the dimensions, \mathbf{Q}_i^l subsumes \mathbf{Q}_j^0 for $2^l(i-1) < j \leq 2^l i$, so that

$$\mathbf{Q}_i^l = \prod_{j=2^l(i-1)+1}^{2^l i} \mathbf{Q}_j^0.$$

At level 0, the algorithm yields the collections \mathbf{Q}_i^0 and \mathbf{P}_i^0 . The remaining clusterings are simply *combinations*, i.e., products of these level-0 clusterings. We denote the collection of $m - 2^l t$ identity matrices (of appropriate size) by \mathbf{I}^l , so that $\mathbf{Q}_1^l = (\mathbf{P}_1^l, \mathbf{I}^l)$. Accoutered with our notation, we now prove the main lemma that relates the combined clustering to its sub-clusterings.

Lemma 3.4. *Let \mathbf{A} be an order- m tensor and $m \geq 2^l t$. The objective function for any $2^l t$ -dimensional clustering $\mathbf{P}_i^l = (\mathbf{P}_{2^l(i-1)+1}^0, \dots, \mathbf{P}_{2^l i}^0)$ can be bound via the sub-clusterings along only one set of dimensions of size t as*

$$\|\mathbf{A} - \mathbf{Q}_i^l \cdot \mathbf{A}\|_F^2 \leq \max_{2^l(i-1) < j \leq 2^l i} 2^l \|\mathbf{A} - \mathbf{Q}_j^0 \cdot \mathbf{A}\|_F^2. \quad (3.4)$$

We can always (wlog) permute dimensions so that any set of 2^l clustered dimensions maps to the first 2^l ones. Hence, it suffices to prove the lemma for $i = 1$, i.e., the first 2^l dimensions.

Proof. We prove the lemma for $i = 1$ by induction on l .

Base: Let $l = 0$. Then $\mathbf{Q}_1^l = \mathbf{Q}_1^0$, and (3.4) holds trivially.

Induction: Assume the claim holds for $l \geq 0$. Consider a clustering $\mathbf{P}_1^{l+1} = (\mathbf{P}_1^l, \mathbf{P}_2^l)$, or equivalently $\mathbf{Q}_1^{l+1} = \mathbf{Q}_1^l \mathbf{Q}_2^l$. Using $\mathbf{P} + \mathbf{P}^\perp = \mathbf{I}$, we decompose \mathbf{A} as

$$\begin{aligned} \mathbf{A} &= (\mathbf{P}_1^{l+1} + \mathbf{P}_1^{l+1\perp}, \mathbf{I}^{l+1}) \cdot \mathbf{A} = (\mathbf{P}_1^l + \mathbf{P}_1^{l\perp}, \mathbf{P}_2^l + \mathbf{P}_2^{l\perp}, \mathbf{I}^{l+1}) \cdot \mathbf{A} \\ &= (\mathbf{P}_1^l, \mathbf{P}_2^l, \mathbf{I}^{l+1}) \cdot \mathbf{A} + (\mathbf{P}_1^{l\perp}, \mathbf{P}_2^l, \mathbf{I}^{l+1}) \cdot \mathbf{A} + (\mathbf{P}_1^l, \mathbf{P}_2^{l\perp}, \mathbf{I}^{l+1}) \cdot \mathbf{A} + (\mathbf{P}_1^{l\perp}, \mathbf{P}_2^{l\perp}, \mathbf{I}^{l+1}) \cdot \mathbf{A} \\ &= \mathbf{Q}_1^l \mathbf{Q}_2^l \cdot \mathbf{A} + \mathbf{Q}_1^{l\perp} \mathbf{Q}_2^l \cdot \mathbf{A} + \mathbf{Q}_1^l \mathbf{Q}_2^{l\perp} \cdot \mathbf{A} + \mathbf{Q}_1^{l\perp} \mathbf{Q}_2^{l\perp} \cdot \mathbf{A}, \end{aligned}$$

where $\mathbf{Q}_1^{l\perp} = (\mathbf{P}_1^{l\perp}, \mathbf{I}^l)$. Since $\mathbf{Q}_1^{l+1} = \mathbf{Q}_1^l \mathbf{Q}_2^l$, the Pythagorean Property 3.2 yields

$$\|\mathbf{A} - \mathbf{Q}_1^{l+1} \cdot \mathbf{A}\|^2 = \|\mathbf{Q}_1^{l\perp} \mathbf{Q}_2^l \cdot \mathbf{A}\|^2 + \|\mathbf{Q}_1^l \mathbf{Q}_2^{l\perp} \cdot \mathbf{A}\|^2 + \|\mathbf{Q}_1^{l\perp} \mathbf{Q}_2^{l\perp} \cdot \mathbf{A}\|^2.$$

Combining the above equalities with the assumption (wlog)

$$\|\mathbf{Q}_1^{l\perp} \mathbf{Q}_2^l \cdot \mathbf{A}\|^2 \geq \|\mathbf{Q}_1^l \mathbf{Q}_2^{l\perp} \cdot \mathbf{A}\|^2,$$

we obtain the inequalities

$$\begin{aligned} \|\mathbf{A} - \mathbf{Q}_1^l \mathbf{Q}_2^l \cdot \mathbf{A}\|^2 &\leq 2(\|\mathbf{Q}_1^{l\perp} \mathbf{Q}_2^l \cdot \mathbf{A}\|^2 + \|\mathbf{Q}_1^{l\perp} \mathbf{Q}_2^{l\perp} \cdot \mathbf{A}\|^2) \\ &= 2\|\mathbf{Q}_1^{l\perp} \mathbf{Q}_2^l \cdot \mathbf{A} + \mathbf{Q}_1^{l\perp} \mathbf{Q}_2^{l\perp} \cdot \mathbf{A}\|^2 = 2\|\mathbf{Q}_1^{l\perp} (\mathbf{Q}_2^l + \mathbf{Q}_2^{l\perp}) \cdot \mathbf{A}\|^2 \\ &= 2\|\mathbf{Q}_1^{l\perp} \cdot \mathbf{A}\|^2 = 2\|\mathbf{A} - \mathbf{Q}_1^l \cdot \mathbf{A}\|^2 \\ &\leq 2 \max_{1 \leq j \leq 2^l} \|\mathbf{A} - \mathbf{Q}_j^l \cdot \mathbf{A}\|^2 \leq 2 \cdot 2^l \max_{1 \leq j \leq 2^{l+1}} \|\mathbf{A} - \mathbf{Q}_j^0 \cdot \mathbf{A}\|^2, \end{aligned}$$

where the last step follows from the induction hypothesis (3.4), and the two norm terms in the first line are combined using the Pythagorean Property. \square

Proof. (Thm. 3.1, Case 1). Let $m = 2^h t$. Using an algorithm with guarantee α_t , we cluster each subset (indexed by i) of t dimensions to obtain \mathbf{Q}_i^0 . Let \mathbf{S}_i be the optimal sub-clustering of subset i , i.e., the result that \mathbf{Q}_i^0 would be if α_t were 1. We bound the objective for the collection of all m sub-clusterings $\mathbf{P}_1^h = \mathbf{Q}_1^h$ as

$$\|\mathbf{A} - \mathbf{Q}_1^h \cdot \mathbf{A}\|_F^2 \leq 2^h \max_j \|\mathbf{A} - \mathbf{Q}_j^0 \cdot \mathbf{A}\|^2 \leq 2^h \alpha_t \max_j \|\mathbf{A} - \mathbf{S}_j \cdot \mathbf{A}\|_F^2. \quad (3.5)$$

The first inequality follows from Lemma 3.4, while the last inequality follows from the α_t approximation factor that we used to get sub-clustering \mathbf{Q}_j^0 .

So far we have related our approximation to an optimal sub-clustering across a set of dimensions. Let us hence look at the relation between such an optimal sub-clustering \mathbf{S} of the first t dimensions (via permutation, these dimensions correspond to an arbitrary subset of size t), and the optimal tensor clustering \mathbf{F} across all the $m = 2^h t$ dimensions. Recall that a clustering can be expressed by either the projection matrices collected in \mathbf{Q}_1^l , or by cluster indicator matrices \mathbf{C}_i together with the mean tensor \mathbf{M} , so that

$$(\mathbf{C}_1, \dots, \mathbf{C}_{2^h t}, \mathbf{I}^l) \cdot \mathbf{M} = \mathbf{Q}_1^l \cdot \mathbf{A}.$$

Let \mathbf{C}_j^S and \mathbf{C}_j^F be the dimension-wise cluster indicator matrices for \mathbf{S} and \mathbf{F} , respectively; By definition, \mathbf{S} solves

$$\min_{\mathbf{C}_1, \dots, \mathbf{C}_t, \mathbf{M}} \|\mathbf{A} - (\mathbf{C}_1, \dots, \mathbf{C}_t, \mathbf{I}^0) \cdot \mathbf{M}\|_{\mathbb{F}}^2, \quad \text{s.t. } \mathbf{C}_j \in \{0, 1\}^{n_j \times k_j},$$

which makes \mathbf{S} even better than the sub-clustering $(\mathbf{C}_1^F, \dots, \mathbf{C}_t^F)$ induced by the optimal m -dimensional clustering \mathbf{F} . Thus,

$$\begin{aligned} \|\mathbf{A} - \mathbf{S} \cdot \mathbf{A}\|_{\mathbb{F}}^2 &\leq \min_{\mathbf{M}} \|\mathbf{A} - (\mathbf{C}_1^F, \dots, \mathbf{C}_t^F, \mathbf{I}^0) \cdot \mathbf{M}\|_{\mathbb{F}}^2 \\ &\leq \|\mathbf{A} - (\mathbf{C}_1^F, \dots, \mathbf{C}_t^F, \mathbf{I}^0)(\mathbf{I}, \dots, \mathbf{I}, \mathbf{C}_{t+1}^F, \dots, \mathbf{C}_m^F) \cdot \mathbf{M}^F\|_{\mathbb{F}}^2 \\ &= \|\mathbf{A} - \mathbf{F} \cdot \mathbf{A}\|_{\mathbb{F}}^2, \end{aligned} \quad (3.6)$$

where \mathbf{M}^F is the tensor of means for the optimal m -dimensional clustering. Combining (3.5) with (3.6) yields the final bound for the combined clustering $\mathcal{C} = \mathbf{Q}_1^h$,

$$J_m(\mathcal{C}) = \|\mathbf{A} - \mathbf{Q}_1^h \cdot \mathbf{A}\|_{\mathbb{F}}^2 \leq 2^h \alpha_t \|\mathbf{A} - \mathbf{F} \cdot \mathbf{A}\|_{\mathbb{F}}^2 = 2^h \alpha_t J_{\text{OPT}}(m),$$

which completes the proof of the theorem. \square

3.3 Analysis: Theorem 3.1, Metric case

Now we present our proof of Thm. 3.1 for the case where $d(x, y)$ is a metric, such as an ℓ_p distance or separable Hilbertian metric. For this case, recall that the tensor clustering problem is

$$\underset{(\mathbf{C}_1, \dots, \mathbf{C}_m), \mathbf{M}}{\text{minimize}} J(\mathcal{C}) = d(\mathbf{A}, (\mathbf{C}_1, \dots, \mathbf{C}_m) \cdot \mathbf{M}), \quad \text{s.t. } \mathbf{C}_j \in \{0, 1\}^{n_j \times k_j}. \quad (3.7)$$

Since in general the best representative \mathbf{M} is not the mean tensor, we cannot use the shorthand $\mathbf{P} \cdot \mathbf{A}$ for \mathbf{M} , so the proof is different from the Euclidean case.

Proof. We will split the dimensions in a different way. Let \mathbf{R}_i^ℓ be the collection of clusterings of dimensions $i, \dots, i + \ell - 1$. \mathbf{R}_i^ℓ combines the \mathcal{C}_j in a manner analogous to how \mathbf{Q}_i^1 combines projection matrices. For simplicity, the proof here is for clustering single dimensions at a time, but it generalizes in a straightforward way to chunks of t dimensions, leading to a factor $2m/t$ instead of $2m$.

Let us first prove a relation for any subset of the last $m - i + 1$ dimensions, $\mathbf{R}_i^1 \mathbf{R}_{i+1}^{m-i} = \mathbf{R}_i^{m-i+1}$. Let $\mathbf{M}_i^\ell = \underset{\mathbf{X}}{\text{argmin}} d(\mathbf{A}, \mathbf{R}_i^\ell \cdot \mathbf{X})$ be the optimal representatives for the clustering collections \mathbf{R}_i^1 and \mathbf{R}_{i+1}^{m-i} , and

$$\hat{\mathbf{M}}_i = \underset{\mathbf{X}}{\text{argmin}} d(\mathbf{R}_i^1 \mathbf{M}_i^1, \mathbf{R}_i^1 \mathbf{R}_{i+1}^{m-i} \cdot \mathbf{X}), \quad \mathbf{X} \in \mathbb{R}^{n_1 \times \dots \times n_{i-1} \times k_i \times \dots \times k_m}.$$

The index ι will run over dimension i , and the multi-indices, r, j over dimensions $1, \dots, i - 1$ and $i + 1, \dots, m$, respectively. The indices I and multi-indices J refer to the clusterings in \mathbf{R}_i^1 and \mathbf{R}_{i+1}^{m-i} , respectively. Since $\hat{\mathbf{M}}_i$ is the element-wise minimum, we have

$$\begin{aligned} d(\mathbf{R}_i^1 \cdot \mathbf{M}_i^1, \mathbf{R}_i^1 \mathbf{R}_{i+1}^{m-i} \cdot \hat{\mathbf{M}}_i) &= \sum_{I, J} \sum_{\iota \in I, r} \min_{\mu_{IJr} \in \mathbb{R}} \sum_{j \in J} d((\mu_\iota^1)_{IJr}, \mu_{IJr}) \\ &\leq \sum_{I, J} \sum_{\iota \in I, r} \sum_{j \in J} d((\mu_\iota^1)_{IJr}, (\mu_{i+1}^{m-i})_{\iota Jr}) = d(\mathbf{R}_i^1 \cdot \mathbf{M}_i^1, \mathbf{R}_{i+1}^{m-i} \cdot \mathbf{M}_{i+1}^{m-1}). \end{aligned}$$

We use this relation and the triangle inequality to break down \mathbf{R}_1^m into its single-dimansional parts.

We then relate the objectives of these parts to the optimal single-dimensional clusterings \mathbf{S}_i^1 .

$$\begin{aligned}
\min_{\mathbf{M}^m} d(\mathbf{A}, \mathbf{R}_1^1 \mathbf{R}_2^{m-1} \cdot \mathbf{M}^m) &\leq d(\mathbf{A}, \mathbf{R}_1^1 \mathbf{R}_2^{m-1} \cdot \hat{\mathbf{M}}_1) \\
&\leq d(\mathbf{A}, \mathbf{R}_1^1 \cdot \mathbf{M}_1^1) + d(\mathbf{R}_1^1 \cdot \mathbf{M}_1^1, \mathbf{R}_1^1 \mathbf{R}_2^{m-1} \cdot \hat{\mathbf{M}}_1) \\
&\leq d(\mathbf{A}, \mathbf{R}_1^1 \cdot \mathbf{M}_1^1) + d(\mathbf{R}_1^1 \cdot \mathbf{M}_1^1, \mathbf{R}_2^{m-1} \cdot \mathbf{M}_2^{m-1}) \\
&\leq 2d(\mathbf{A}, \mathbf{R}_1^1 \cdot \mathbf{M}_1^1) + d(\mathbf{A}, \mathbf{R}_2^{m-1} \cdot \mathbf{M}_2^{m-1}) \\
&\leq 2d(\mathbf{A}, \mathbf{R}_1^1 \cdot \mathbf{M}_1^1) + 2d(\mathbf{A}, \mathbf{R}_2^1 \cdot \mathbf{M}_2^1) + d(\mathbf{A}, \mathbf{R}_2^{m-1} \cdot \mathbf{M}_2^{m-1}) \\
&\leq \dots
\end{aligned} \tag{3.8}$$

$$\leq 2 \sum_{i=1}^m d(\mathbf{A}, \mathbf{R}_i^1 \cdot \mathbf{M}_i^1) \leq 2 \sum_{i=1}^m \alpha_1 \min_{\mathbf{X}} d(\mathbf{A}, \mathbf{S}_i^1 \cdot \mathbf{X}^1). \tag{3.9}$$

For (3.8), we applied the same steps as before to \mathbf{R}_2^1 and \mathbf{R}_3^{m-2} , and then continued this breakdown, always splitting off the first dimension. The last relation follows from the 1D approximation algorithm that was used. What is left is to bound (3.9) by the objective for the optimal m -dimensional clustering $\mathbf{F} \cdot \mathbf{M}_F = \mathbf{F}_1 \mathbf{F}_2 \dots \mathbf{F}_m \cdot \mathbf{M}_F$. Note that, since non-clustered dimensions have identity matrices, the cluster parts commute: $\mathbf{F}_i \mathbf{F}_j \mathbf{X} = \mathbf{F}_j \mathbf{F}_i \mathbf{X}$. Owing to the optimality of \mathbf{S}_i^1 , we have

$$\min_{\mathbf{X}^1} d(\mathbf{A}, \mathbf{S}_i^1 \cdot \mathbf{X}^1) \leq \min_{\mathbf{Y}^1} d(\mathbf{A}, \mathbf{F}_i \cdot \mathbf{Y}^1) \leq \min_{\mathbf{Y}^m} d(\mathbf{A}, \mathbf{F}_i^1 (\mathbf{F}_1^1 \dots \mathbf{F}_{i-1}^1 \mathbf{F}_{i+1}^1 \dots \mathbf{F}_m^1 \cdot \mathbf{Y}^m)) = d(\mathbf{A}, \mathbf{F} \cdot \mathbf{M}_F)$$

for any term in the sum (3.9). Thus, it follows that

$$\min_{\mathbf{M}^m} d(\mathbf{A}, \mathbf{R}_1^m \cdot \mathbf{M}^m) \leq 2 \sum_{i=1}^m \alpha_1 \min_{\mathbf{X}} d(\mathbf{A}, \mathbf{S}_i^1 \cdot \mathbf{X}^1) \leq 2m\alpha_1 d(\mathbf{A}, \mathbf{F} \cdot \mathbf{M}_F),$$

which completes the proof. \square

3.4 Theorem 3.1 with Bregman divergences

Theorem 3.1 also applies to Bregman divergences, i.e., divergences that can be bounded in terms of squared Euclidean distances and for which the best representative is the tensor of means defined in Equation (3.2) [7].

The *Bregman divergence* $B_f(x, y)$ between scalars x and y is defined as [12, 13]

$$B_f(x, y) = f(x) - f(y) - f'(y)(x - y), \tag{3.10}$$

for a given strictly convex function f . With $f = \frac{1}{2}x^2$ the divergence (3.10) reduces to the familiar Euclidean distance $\frac{1}{2}(x - y)^2$, while for $f(x) = x \log x$ it turns into the (generalized) KL Divergence. For tensors, we extend Definition (3.10) by considering *separable* Bregman divergences, so that

$$B_f(\mathbf{X}, \mathbf{Y}) = \sum_{i_1, \dots, i_m} B_f(x_{i_1 \dots i_m}, y_{i_1 \dots i_m}).$$

Let σ_U and σ_L be upper and lower bounds, respectively, with $\sigma_L > 0$, such that

$$\sigma_L B_f(x, y) \leq \|x - y\|^2 \leq \sigma_U B_f(x, y) \tag{3.11}$$

for all x, y in the convex hull of the entries of the given tensor \mathbf{A} . For KL-divergence, the data must then be bounded away from zero.

Since the means tensor is the best representative $\arg\min_{\mathbf{X}} B_f(\mathbf{A}, \mathbf{R} \cdot \mathbf{A})$ for a clustering \mathbf{R} , we again use projection matrices to express clusterings. Let \mathbf{Q}_1^h be, as above, the full combination of projection matrices from dimension-wise clustering, and $\mathbf{F} = \arg\min_{\mathbf{Q}} B_f(\mathbf{A}, \mathbf{Q} \cdot \mathbf{A})$ the optimal

m -dimensional tensor clustering. Then we know that

$$\begin{aligned} B_f(\mathbf{A}, \mathbf{Q}_1^h) &\leq \sigma_U \|\mathbf{A}, \mathbf{Q}_1^h\|^2 \\ &\leq \sigma_U 2^{\log_2 m/t} \max_j \|\mathbf{A} - \mathbf{Q}_j^0 \cdot \mathbf{A}\|^2 \end{aligned} \quad (3.12)$$

$$\begin{aligned} &\leq \frac{\sigma_U}{\sigma_L} 2^{\log_2 m/t} \max_j D(\mathbf{A}, \mathbf{Q}_j^0 \cdot \mathbf{A}) \\ &\leq \frac{\sigma_U}{\sigma_L} 2^{\log_2 m/t} B_f(\mathbf{A}, \mathbf{F} \cdot \mathbf{A}), \end{aligned} \quad (3.13)$$

so $\rho_d = \frac{\sigma_U}{\sigma_L}$. Inequality (3.12) follows from Lemma 3.4, and Inequality (3.13) from an argumentation analogous to Equation (3.6).

Curvature bounds as in (3.11) seem to be necessary for Bregman divergences to guarantee *constant* approximation factors for the underlying 1D clustering—this intuition is reinforced by the results of [14], who avoided such curvature assumptions and had to be content with a *non-constant* $O(\log n)$ approximation factor for information theoretic clustering.

3.5 Implications

To obtain concrete bounds for a variety of tensor clustering problems, we can use Theorem 3.1 for $t = 1$ or $t = 2$ with existing 1D approximation factors α_t from the literature. Table 1 summarizes the results.

3.5.1 1D factors for Metric and Bregman clustering

The $(1 + \epsilon)$ approximation factor for 1D clustering by Ackermann et al. [2] applies to all metrics. It leads to an m -dimensional approximation factor of $\alpha_m = p(m/t)(1 + \epsilon)$. Arthur and Vassilvskii [6] prove a guarantee in expectation of $\alpha_1 = 8(\log K + 2)$ for K clusters with Euclidean k-means, resulting in an expected $\alpha_m = 8p(m/t)(\log K + 2)$.

For Bregman clustering, we arrive at similar results with the approximation factor by Ackermann and Blömer [1] or the extension of [6] in [30, 35].

3.5.2 Hilbertian metrics

A special example of metrics are Hilbertian metrics [24, 33] that arise from conditionally positive definite (CPD) kernels. A real valued function $C : \mathcal{S} \times \mathcal{S} \mapsto \mathbb{R}$ is called a *conditionally positive definite* (CPD) kernel on \mathcal{S} if for any positive integer n , any choice of n elements $x_i \in \mathcal{S}$, $[i]_1^n$ ($[i]_1^n \equiv i = 1, \dots, n$) and any choice of n reals $u_i \in \mathbb{R}$ such that $\sum_i u_i = 0$, we have $\sum_{i,j=0}^n u_i u_j C(x_i, x_j) \geq 0$ [11, 33]. The following remarkable result [32] connects CPD kernels and Hilbertian metrics, i.e., metrics which can be isometrically embedded in Hilbert space: There exists a Hilbert space \mathcal{H} of real-valued functions on \mathcal{S} , and a mapping $\Phi : \mathcal{S} \mapsto \mathcal{H}$ such that

$$\|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\|^2 = -C(\mathbf{x}, \mathbf{y}) + \frac{1}{2}(C(\mathbf{x}, \mathbf{x}) + C(\mathbf{y}, \mathbf{y})) = d_C(\mathbf{x}, \mathbf{y}),$$

if and only if $C(\cdot, \cdot)$ is a CPD kernel. Hence, given a CPD kernel C , one can construct a Hilbertian metric $d_C(\mathbf{x}, \mathbf{y})$ which behaves like the squared Euclidean distance in the Hilbert space. The corresponding kernel is $K(x, y) = \frac{1}{2}(C(x, y) - C(x, a) - C(y, a) + C(a, a))$ for some fixed $a \in \mathcal{S}$.

Here, we choose $\mathcal{S} \subseteq \mathbb{R}$ and define the distance of tensors \mathbf{X}, \mathbf{Y} as

$$d_C(\mathbf{X}, \mathbf{Y}) = \sum_{i_1, \dots, i_m} d_C(x_{i_1, \dots, i_m}, y_{i_1, \dots, i_m}).$$

Since the argument by [6] for their kmeans++ is independent of the dimensionality, it can be generalized from Euclidean distance to distances in a Hilbert space.

Lemma 3.5 (1D Hilbertian Metric Clustering). *For any 1D clustering with a Hilbertian metric d_C , one can construct a `kmeans++` based initialization followed by iterative updates using kernel `k-means` such that if \mathcal{C} is the final clustering, then*

$$E[J(\mathcal{C})] \leq 8(\log K + 2)J_{OPT} . \quad (3.14)$$

Proof. Using $d_C(x, y) = \|\Phi(x) - \Phi(y)\|^2$, we can use the initialization by [6] in the Hilbert space on the mapped data points $\Phi(x)$, since it only depends on squared Euclidean distances or inner products, independent of the dimensionality of the space. Finally, the objective function can always be improved by running kernel `kmeans` starting from the `kmeans++` initialization. \square

Together with Theorem 3.1, Lemma 3.5 directly leads to a tensor clustering guarantee for Hilbertian metrics:

$$E[J(\mathcal{C})] \leq 8m(\log K^* + 2)J_{OPT}(m), \quad (3.15)$$

where $K^* = \max_{1 \leq j \leq m} k_j$ is the maximum number of clusters across all dimensions.

3.5.3 2D factor for binary ℓ_1 clustering

Applying the results of [31] for binary matrices as α_2 yields the slightly stronger bound for ℓ_1 tensor clustering:

$$J(\mathcal{C}) \leq 3^{\log_2(m)-1}(1 + \sqrt{2})\alpha_1 J_{OPT}(m).$$

Table 1: Approximation guarantees for Tensor Clustering Algorithms. K^* denotes the maximum number of clusters, i.e., $K^* = \arg\max_j k_j$; c is some constant.

Problem Name	Approx. Bound	Proof
Metric tensor clustering	$J(\mathcal{C}) \leq m(1 + \epsilon)J_{OPT}(m)$	Thm. 3.1 + [2]
Bregman tensor clustering	$E[J(\mathcal{C})] \leq 8mc(\log K^* + 2)J_{OPT}(m)$	(3.11), Thm. 3.1 + [30, 35] (using [6])
Bregman tensor clustering	$J(\mathcal{C}) \leq m\sigma_U\sigma_L^{-1}(1 + \epsilon)J_{OPT}(m)$	(3.11), Thm. 3.1 + [1]
Bregman co-clustering	Above two results with $m = 2$	as above
Hilbertian metrics	$E[J(\mathcal{C})] \leq 8m(\log K^* + 2)J_{OPT}(m)$	Thm. 3.1 + Lemma 3.5

4 Experiments

Our bounds depend strongly on the approximation factor α_t of an underlying t -dimensional clustering method. In our experiments, we study this close dependence for $t = 1$, wherein we compare the tensor clusterings arising from different 1D methods of varying sophistication. Keep in mind that the comparison of the 1D methods is to see their impact on the tensor clustering built on top of them.

Our experiments reveal that the empirical approximation factors are usually smaller than the theoretical bounds, and these factors depend on statistical properties of the data. We also observe the linear dependence of the CoTeC objectives on the associated 1D objectives, as suggested by Thm. 3.1 (for Euclidean) and Table 1 (2nd row, for KL-Divergence).

Further comparisons show that in practice, CoTeC is competitive with a greedy heuristic SiTeC (**S**imultaneous **T**ensor **C**lustering), which *simultaneously* takes *all* dimensions into account, but lacks theoretical guarantees. As expected, initializing SiTeC with CoTeC yields lower final objective values using fewer “simultaneous” iterations.

Regarding divergences, we focus on Euclidean distance and KL-divergence to test CoTeC. To study the effect of the 1D method, we use two seeding methods for each divergence, uniform and distance-based drawing. The latter seeding ensures 1D approximation factors for $E[J(\mathcal{C})]$ by [6] for Euclidean clustering and by [30, 35] for KL-divergence.

We use each seeding by itself and as an initialization for `k-means` to get four 1D methods for each divergence. We refer to the CoTeC combination of the corresponding independent 1D clusterings by abbreviations:

r: Randomly (uniformly) sample centers from the data points; assign each point to its closest center.
s: Sample centers using distance-specific seeding [6, 30, 35]; assign each point to its closest center.
rk: Initialize Euclidean or Bregman k-means with ‘**r**’.
sk: Initialize Euclidean or Bregman k-means with ‘**s**’.

The SiTeC method we compare to is the minimum sum-squared residue co-clustering of [16] for Euclidean distances in 2D, and a generalization of Algorithm 1 of [9] for 3D and Bregman 2D clustering. Additionally, we initialize SiTeC with the outcome of each of the four CoTeC variants, which yields four versions (of SiTeC), namely,

We compare the four versions of CoTeC to SiTeC, an algorithm without guarantees that considers the groupings in all dimensions together. For Euclidean distances in 2D, we use the minimum sum-squared residue co-clustering of [16] as SiTeC, while for Euclidean 3D and Bregman tensor clustering, we generalize Algorithm 1 of [9]. Initializing SiTeC with each one of the above schemes results in another four variants:

rc: SiTeC initialized with the results of ‘**r**’
sc: SiTeC initialized with the results of ‘**s**’
rk: SiTeC initialized with the results of ‘**rk**’
sk: SiTeC initialized with the results of ‘**sk**’

These variants inherit the guarantees of CoTeC, as they monotonically decrease the objective value.

4.1 Experiments on synthetic data

For a controlled setting with synthetic data, we generate tensors **A** of size $75 \times 75 \times 50$ and 75×75 , for which we randomly choose a $5 \times 5 \times 5$ tensor of means **M** and cluster indicator matrices $C_i \in \{0, 1\}^{n_i \times 5}$. For clustering with Euclidean distances we add Gaussian noise (from $\mathcal{N}(0, \sigma^2)$ with varying σ) to **A**, while for KL-Divergences we use the sampling method of [9] with varying noise.

For each noise-level to test, we repeat the 1D seeding 20 times on each of five generated tensors and average the resulting 100 objective values. To estimate the approximation factor α_m on a tensor, we divide the achieved objective $J(\mathcal{C})$ by the objective value of the “true” underlying tensor clustering. Figure 1 shows the empirical approximation factor $\hat{\alpha}_m$ for Euclidean distance and KL-Divergence. Qualitatively, the plots for tensors of order 2 and 3 do not differ.

In all settings, the empirical factor remains below the theoretical factor. The reason for decreasing approximation factors with higher noise could be lower accuracy of the estimates of $J(\mathcal{C})$ on the one hand, and more similar objective values for all clusterings on the other hand. With low noise, distance-specific seeding **s** yields better results than uniform seeding **r**, and adding k-means on top (**rk**, **sk**) improves the results of both. With Euclidean distances, CoTeC with well-initialized 1D k -means (**sk**) competes with SiTeC. For KL-divergence, though, SiTeC still improves on **sk**, and with high noise levels, 1D k -means does not help: both **rk** and **sk** are as good as their seeding only counterparts.

In summary, the empirical approximation factor does depend on the data, but in general seems to be lower than the theoretical worst-case value.

4.2 Experiments on real data

We further assess the behavior of CoTeC on a number of real-world gene expression data sets⁵.

The first three of our data sets, *Bcell* (1332×62), *AllAml* (2088×72) and *Breast* (21906×77) are gene expression microarray data sets, and described in detail in [25]. *Bcell* is a lymphoma microarray dataset of chronic lymphocytic leukemia, diffuse large Bcell leukemia and follicular lymphoma. During preprocessing only those genes were selected whose minimum expression level was above e^{-1000} . Microarray data for B-cell and T-cell acute lymphocytic leukemia and acute myelogenous leukemia is collected in *AllAml*. Our data matrix is restricted to those genes whose ratio of maximum to

⁵We thank Hyuk Cho for kindly providing us the preprocessed data.

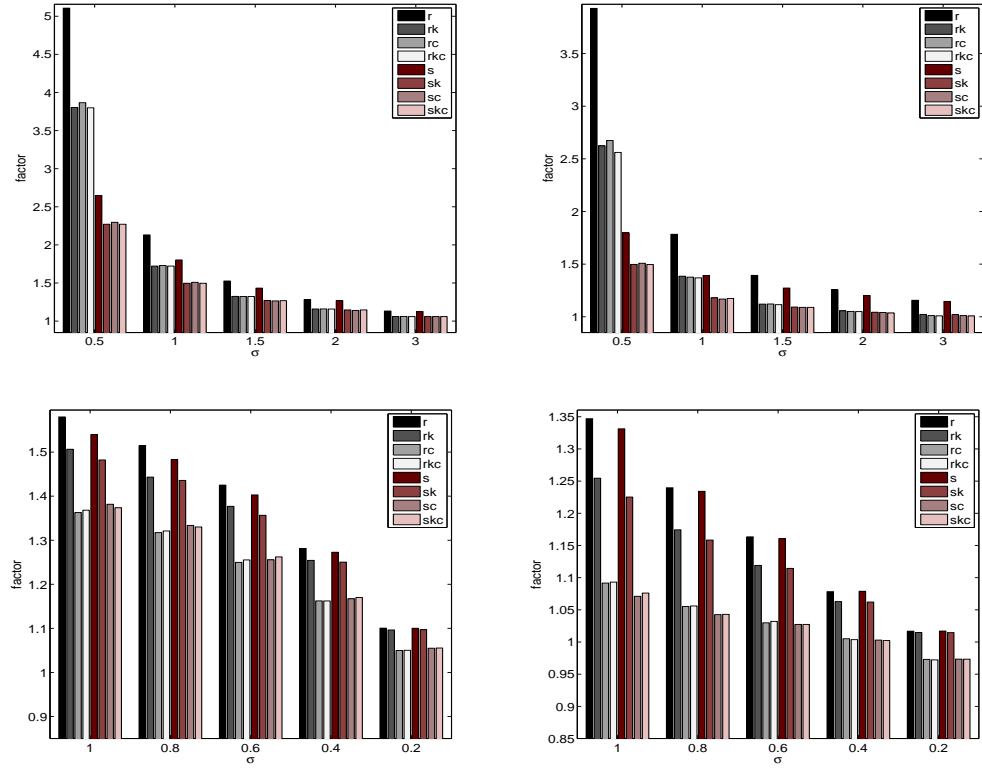


Figure 1: Approximation factors for 3D clustering (left) and co-clustering (right) with increasing noise. Top row: Euclidean distances, bottom row: KL Divergence. The x axis shows σ , the y axis the empirical approximation factor.

minimum expression exceeds 10 and for whom the difference between maximum and minimum expression was at least 1000. *Breast* refers to breast cancer data. The gene selection was the same as for *Bcell*.

The remaining two data sets are cancer microarray matrices from [15]. *Leukemia* (3571×72) [20] is data from acute lymphoblastic leukemia or acute myeloid leukemia, and *Mll* (2474×72) [5] includes data from three types of leukemia (ALL, AML, MLL).

Even though the data sets have labeled column clusters, we do not compare clustering results with the true labels, as the algorithm and its guarantees hold merely for the clustering objective function, which may not exactly agree with the true labels. Moreover, we aim for a co-clustering result and not single-dimensional clusterings, and the labels are available for only one of the dimensions.

For each data set, we repeat the sampling of centers 30 times and average the resulting objective values. Tables 2 to 4 show detailed results. Panel (i) displays the objective value for the simplest CoTeC, **r**, as a baseline, and the relative improvement achieved by the other methods. The methods are encoded as **x**, **xk**, **xc**, **xkc**, where **x** stands for **r** or **s**, depending on the row in the said table.

Overall, the improvements obtained via the approximation algorithm do depend on the dataset under consideration and the number of clusters sought. In general, the improvements are lower for the bispherically normalized data (e.g., that of [15]) than for the other data sets.

For both distances, using 1D k-means on top of the seeding generally improves on the combined co-clustering. The combination method seems particularly competitive for Euclidean distances. On the *Bcell* data (Table 2), the **s** variant of CoTeC (without k-means) can be *as good as* SiTeC **r** initialization. The distance-specific seeding (**s**) gains compared to uniform seeding as the clusters become smaller. For *Bcell* and *Breast* (Table 2), the combination of 1D k-means clusterings (**rk** and **sk**) slightly outperforms the SiTeC variants **rc** and **sc**.

Turning to KL Divergences, the impact of the 1D method varies with the data, as for Euclidean distance. Both 1D k-means and better seeding mostly improve the overall outcome. We observe the highest improvements on the *AllAml* data set. With KL Divergences, SiTeC is almost always at least a bit better than CoTeC.

Besides improving the final result, a good initialization aids SiTeC in yet another way: the average number of iterations it takes to converge decreases, at times to even less than half the reference value.

Overall, the experiments demonstrate that the combination of good single-dimensional clusterings can already lead to reasonable co-clusterings in practice, which can at times be as good as the result of a simultaneous biclustering method. Used as an initialization, the CoTeC results improve the outcome of SiTeC and reduce the number of “simultaneous” iterations.

5 Conclusions

In this paper we presented a simple, and to our knowledge the first approximation algorithm for Bregman and metric tensor clustering. Our approximation factor grows linearly with the order m of the tensor for Bregman divergences, and is slightly superlinear in m for arbitrary metrics. It is always linear in the quality of the sub-clusterings.

Our experiments demonstrated the dependence of the multi-dimensional clustering on the single-dimensional clusterings, confirming the dependence stated in the theoretical bound. On real-world data, the approximation algorithm is also suitable as an initialization for a simultaneous co-clustering algorithm, and endows the latter with its approximation guarantees. In fact the approximation algorithm by itself can also yield reasonable results in practice.

In our experiments we used single-dimensional clusterings with guarantees for our overall approximation algorithm. An interesting direction for future work is the development of a simultaneous approximation algorithm, such as a specific co-clustering seeding scheme of multi-dimensional centers, which can be then used as a subroutine by our tensor clustering algorithm.

Table 2: (i) Improvement of CoTeC and SiTeC variants upon ‘r’ in %; the respective reference value (J_2 for ‘r’) is shaded in gray. (ii) Average number of SiTeC iterations.

Bcell, Euc.						Bcell, KL									
(i)	k_1	k_2	x	xk	xc	xkc	(i)	k_1	k_2	x	xk	xc	xkc		
	5	3	r	$6.00 \cdot 10^5$	20.98	18.37	26.44		5	3	r	$3.73 \cdot 10^{-1}$	15.01	20.87	21.13
			s	8.52	24.97	22.83	29.53				s	1.53	14.31	20.43	20.26
	5	6	r	$5.94 \cdot 10^5$	30.68	26.09	34.72		5	6	r	$3.60 \cdot 10^{-1}$	15.76	21.23	21.62
			s	16.97	33.35	32.06	37.33				s	3.24	16.22	21.37	21.21
	20	3	r	$5.75 \cdot 10^5$	31.66	20.05	33.05		20	3	r	$3.37 \cdot 10^{-1}$	17.59	22.23	23.26
			s	18.83	32.24	24.61	33.36				s	10.54	18.44	22.99	22.98
	20	6	r	$5.56 \cdot 10^5$	49.13	35.26	50.37		20	6	r	$3.15 \cdot 10^{-1}$	18.62	24.51	25.43
			s	34.97	50.55	43.93	51.66				s	11.76	20.52	25.69	26.23
	50	3	r	$5.63 \cdot 10^5$	31.10	14.77	31.76		50	3	r	$3.20 \cdot 10^{-1}$	15.70	20.12	21.07
			s	15.25	32.58	19.14	33.17				s	9.61	17.24	20.85	21.33
	50	6	r	$5.18 \cdot 10^5$	47.55	34.63	48.41		50	6	r	$2.85 \cdot 10^{-1}$	16.38	21.61	22.57
			s	36.22	49.83	43.77	50.55				s	11.86	18.63	23.24	23.13

(ii) k_1 k_2						(ii) k_1 k_2							
			rc	rkc	sc	skc				rc	rkc	sc	skc
5	3		11.9 ± 3.3	3.3 ± 0.7	6.1 ± 2.8	3.5 ± 0.7	5	3		10.1 ± 3.0	7.2 ± 3.0	11.1 ± 4.3	7.2 ± 3.5
5	6		11.9 ± 2.6	3.7 ± 1.7	6.6 ± 2.4	3.3 ± 1.3	5	6		10.8 ± 3.1	8.1 ± 3.4	8.7 ± 2.9	6.8 ± 3.3
20	3		7.0 ± 1.4	2.0 ± 0.2	3.9 ± 1.0	2.2 ± 0.5	20	3		10.6 ± 2.8	7.5 ± 2.0	7.4 ± 1.8	7.0 ± 2.2
20	6		11.3 ± 2.3	2.6 ± 0.8	5.1 ± 2.0	2.7 ± 0.7	20	6		12.6 ± 3.4	8.8 ± 2.9	8.4 ± 2.1	8.1 ± 2.0
50	3		6.2 ± 1.9	2.0 ± 0.0	3.5 ± 2.0	2.0 ± 0.0	50	3		9.1 ± 2.3	6.2 ± 1.3	6.9 ± 1.8	6.0 ± 1.3
50	6		8.1 ± 2.1	2.1 ± 0.3	4.1 ± 1.6	2.0 ± 0.0	50	6		10.5 ± 1.8	7.7 ± 2.1	8.1 ± 2.3	6.9 ± 1.0

Breast, Euc						Breast, KL									
(i)	k_1	k_2	x	xk	xc	xkc	(i)	k_1	k_2	x	xk	xc	xkc		
	5	2	r	$1.43 \cdot 10^5$	22.96	20.48	24.47		5	2	r	$2.70 \cdot 10^{-2}$	8.08	12.81	12.23
			s	2.69	21.92	19.42	24.32				s	1.77	7.98	13.19	12.38
	5	4	r	$1.42 \cdot 10^5$	26.49	25.85	27.30		5	4	r	$2.67 \cdot 10^{-2}$	11.88	17.56	17.31
			s	10.38	26.72	26.67	27.95				s	3.60	11.95	18.10	18.29
	10	2	r	$1.41 \cdot 10^5$	22.13	15.46	25.26		10	2	r	$2.66 \cdot 10^{-2}$	8.01	11.44	12.37
			s	7.77	21.66	19.20	25.09				s	2.45	7.96	12.34	12.46
	10	4	r	$1.37 \cdot 10^5$	26.36	24.09	28.93		10	4	r	$2.59 \cdot 10^{-2}$	11.17	16.54	17.92
			s	9.79	26.87	26.44	29.90				s	4.97	13.53	19.50	19.31
	20	2	r	$1.41 \cdot 10^5$	22.46	10.42	26.21		20	2	r	$2.63 \cdot 10^{-2}$	6.27	9.72	9.95
			s	8.16	22.54	19.43	26.16				s	2.93	8.78	11.69	11.61
	20	4	r	$1.37 \cdot 10^5$	27.95	23.44	31.71		20	4	r	$2.56 \cdot 10^{-2}$	11.73	17.42	17.78
			s	10.55	28.31	25.83	32.44				s	3.45	12.21	17.51	17.45

(ii) k_1 k_2						(ii) k_1 k_2							
			rc	rkc	sc	skc				rc	rkc	sc	skc
5	2		4.6 ± 2.4	1.2 ± 0.4	4.0 ± 1.6	1.8 ± 0.4	5	2		5.2 ± 2.0	3.6 ± 2.0	4.9 ± 2.6	3.1 ± 1.8
5	4		4.9 ± 1.8	1.0 ± 0.2	3.0 ± 0.9	1.2 ± 0.5	5	4		5.6 ± 1.8	3.6 ± 1.9	4.4 ± 1.2	3.5 ± 1.4
10	2		3.4 ± 1.4	2.0 ± 0.2	2.6 ± 1.0	2.0 ± 0.0	10	2		4.0 ± 1.8	2.5 ± 1.0	4.4 ± 2.8	2.7 ± 1.7
10	4		4.3 ± 1.8	2.0 ± 0.5	3.0 ± 0.9	2.1 ± 0.3	10	4		5.1 ± 1.4	4.0 ± 1.7	5.2 ± 1.7	3.7 ± 1.3
20	2		2.9 ± 1.3	2.0 ± 0.0	2.7 ± 1.0	2.0 ± 0.0	20	2		3.6 ± 1.8	2.3 ± 0.9	3.2 ± 1.5	2.1 ± 0.5
20	4		3.9 ± 1.3	2.1 ± 0.3	3.4 ± 1.8	2.0 ± 0.2	20	4		5.2 ± 1.9	3.5 ± 1.8	4.3 ± 1.6	2.8 ± 1.2

Table 3: (i) Improvement of CoTeC and SiTeC variants upon ‘r’ in %; the respective reference value (J_2 for ‘r’) is shaded in gray. (ii) Average number of SiTeC iterations.

AllAml, Euc.						AllAml, KL									
(i)	k_1	k_2	x	xk	xc	xkc	(i)	k_1	k_2	x	xk	xc	xkc		
	5	3	r	$6.06 \cdot 10^{11}$	49.26	49.19	50.54		5	3	r	$5.64 \cdot 10^{-1}$	43.92	47.14	46.73
			s	40.63	48.62	50.27	50.71				s	33.39	43.12	46.68	46.44
	10	3	r	$5.31 \cdot 10^{11}$	47.02	47.01	48.69		10	3	r	$4.67 \cdot 10^{-1}$	40.11	41.72	42.57
			s	40.83	48.65	49.51	50.10				s	31.04	39.87	42.78	42.75
	20	3	r	$4.37 \cdot 10^{11}$	39.75	38.02	41.78		20	3	r	$3.78 \cdot 10^{-1}$	29.29	32.67	33.24
			s	34.26	41.06	42.70	43.28				s	20.58	29.74	33.90	34.07
(ii)	k_1	k_2	rc	rkc	sc	skc	(ii)	k_1	k_2	rc	rkc	sc	skc		
	5	3	13.8 ± 3.7	2.8 ± 1.2	5.2 ± 1.7	3.0 ± 1.4		5	3	17.9 ± 3.5	7.0 ± 3.5	11.7 ± 5.0	7.8 ± 4.1		
	10	3	15.9 ± 4.6	3.4 ± 1.3	4.8 ± 1.2	2.9 ± 1.0		10	3	18.3 ± 3.4	7.2 ± 2.6	12.1 ± 3.5	9.3 ± 4.6		
	20	3	12.3 ± 3.4	2.9 ± 1.3	4.7 ± 1.3	2.9 ± 0.9		20	3	18.9 ± 2.5	12.0 ± 4.5	11.1 ± 3.1	10.3 ± 2.9		
Leukemia, Euc.						Leukemia, KL									
(i)	k_1	k_2	x	xk	xc	xkc	(i)	k_1	k_2	x	xk	xc	xkc		
	3	2	r	$7.61 \cdot 10^4$	5.48	5.77	6.74		3	2	r	$1.82 \cdot 10^{-1}$	5.11	7.15	7.52
			s	0.17	5.54	5.73	6.78				s	0.36	4.93	7.19	7.51
	3	3	r	$7.57 \cdot 10^4$	6.53	7.18	7.75		3	3	r	$1.81 \cdot 10^{-1}$	6.00	8.13	8.76
			s	0.14	6.79	6.77	7.79				s	0.44	6.08	8.18	8.76
	50	2	r	$7.30 \cdot 10^4$	3.79	5.97	7.25		50	2	r	$1.71 \cdot 10^{-1}$	3.81	7.58	7.60
			s	0.33	3.75	5.54	7.25				s	-0.21	3.65	7.32	7.35
	50	3	r	$7.15 \cdot 10^4$	4.90	7.34	8.93		50	3	r	$1.68 \cdot 10^{-1}$	4.74	9.31	9.35
			s	0.60	5.00	8.00	9.06				s	1.08	5.16	9.70	9.75
	75	2	r	$7.26 \cdot 10^4$	3.66	5.67	6.89		75	2	r	$1.71 \cdot 10^{-1}$	3.36	6.92	6.95
			s	0.02	3.67	5.23	6.88				s	-0.35	2.85	6.60	6.30
	75	3	r	$7.09 \cdot 10^4$	4.59	7.09	8.47		75	3	r	$1.66 \cdot 10^{-1}$	4.48	9.04	9.11
			s	0.60	4.61	7.05	8.52				s	0.69	4.25	8.66	8.68
(ii)	k_1	k_2	rc	rkc	sc	skc	(ii)	k_1	k_2	rc	rkc	sc	skc		
	3	2	3.8 ± 1.3	2.0 ± 0.0	3.3 ± 0.8	2.0 ± 0.0		3	2	7.6 ± 3.5	4.5 ± 3.2	8.0 ± 2.9	4.6 ± 3.2		
	3	3	4.5 ± 1.5	2.2 ± 0.4	3.8 ± 1.1	2.1 ± 0.3		3	3	7.4 ± 2.5	5.1 ± 1.7	7.3 ± 3.0	4.7 ± 1.4		
	50	2	3.3 ± 1.1	2.0 ± 0.0	2.9 ± 1.3	2.0 ± 0.0		50	2	5.4 ± 1.8	3.4 ± 0.7	5.7 ± 2.5	3.3 ± 0.5		
	50	3	3.3 ± 0.8	2.0 ± 0.0	3.7 ± 1.1	2.0 ± 0.0		50	3	6.2 ± 2.0	4.5 ± 0.8	5.5 ± 1.0	4.6 ± 1.1		
	75	2	3.1 ± 0.9	2.0 ± 0.0	3.3 ± 1.1	2.0 ± 0.0		75	2	5.3 ± 1.8	3.4 ± 1.2	5.6 ± 2.2	3.2 ± 0.5		
	75	3	3.6 ± 0.9	2.0 ± 0.0	3.4 ± 1.0	2.0 ± 0.0		75	3	5.6 ± 1.4	4.2 ± 0.6	4.9 ± 1.1	4.1 ± 0.3		

Table 4: (i) Improvement of CoTeC and SiTeC variants upon ‘r’ in %; the respective reference value (J_2 for ‘r’) is shaded in gray. (ii) Average number of SiTeC iterations.

Mil, Euc.						
(i)	k_1	k_2	x	xk	xc	xkc
	3	3 r	$6.52 \cdot 10^4$	10.54	11.26	11.45
		s	1.41	10.62	11.20	11.46
50	3 r		$5.83 \cdot 10^4$	8.40	12.53	13.21
		s	1.12	8.23	12.35	13.17
75	3 r		$5.75 \cdot 10^4$	7.84	11.69	12.52
		s	0.84	7.86	11.68	12.52

(ii)	k_1	k_2	rc	rkc	sc	skc
	3	3	4.2 ± 1.2	2.0 ± 0.3	3.8 ± 1.0	2.0 ± 0.5
	50	3	4.7 ± 1.9	2.0 ± 0.0	4.2 ± 1.5	2.1 ± 0.3
75	3		4.4 ± 1.4	2.0 ± 0.0	4.3 ± 1.5	2.0 ± 0.0

References

- [1] M. R. Ackermann and Johannes Blömer. Coresets and Approximate Clustering for Bregman Divergences. In *Proc. 20th ACM-SIAM Symposium on Discrete Algorithms (SODA '09)*, 2009. To appear.
- [2] M. R. Ackermann, J. Blömer, and C. Sohler. Clustering for metric and non-metric distance measures. In *ACM-SIAM SODA*, April 2008.
- [3] S. Agarwal, J. Lim, L. Zelnik-Manor, P. Perona, D. Kriegman, and S. Belongie. Beyond pairwise clustering. In *IEEE CVPR*, 2005.
- [4] A. Anagnostopoulos, A. Dasgupta, and R. Kumar. Approximation algorithms for co-clustering. In *PODS*, 2008.
- [5] S. A. Armstrong. M1 translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30:41–17, 2002.
- [6] D. Arthur and S. Vassilvitskii. **k-means++**: The Advantages of Careful Seeding. In *SODA*, pages 1027–1035, 2007.
- [7] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman Divergences. *JMLR*, 6(6):1705–1749, October 2005.
- [8] A. Banerjee, S. Basu, and S. Merugu. Multi-way Clustering on Relation Graphs. In *SIAM Data Mining*, 2007.
- [9] A. Banerjee, I. S. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha. A Generalized Maximum Entropy Approach to Bregman Co-clustering and Matrix Approximation. *JMLR*, 8:1919–1986, 2007.
- [10] R. Bekkerman, R. El-Yaniv, and A. McCallum. Multi-way distributional clustering via pairwise interactions. In *ICML*, 2005.
- [11] C. Berg, J. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer-Verlag, 1984.
- [12] L. M. Bregman. The relaxation method of finding the common point of convex sets and its applications to the solution of problems in convex programming. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.
- [13] Y. Censor and S. A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, 1997.
- [14] K. Chaudhuri and A. McGregor. Finding metric structure in information theoretic clustering. In *Conf. on Learning Theory, COLT*, July 2008.
- [15] H. Cho and I. Dhillon. Coclustering of human cancer microarrays using minimum sum-squared residue coclustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(3):385–400, 2008.
- [16] H. Cho, I. S. Dhillon, Y. Guan, and S. Sra. Minimum Sum Squared Residue based Co-clustering of Gene Expression data. In *Proc. 4th SIAM International Conference on Data Mining (SDM)*, pages 114–125, Florida, 2004. SIAM.
- [17] V. de Silva and L.-H. Lim. Tensor Rank and the Ill-Posedness of the Best Low-Rank Approximation Problem. *SIAM J. on Matrix Analysis and Applications*, 30(3):1084–1127, 2008.
- [18] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *Proc. ACM SIGKDD 2003*, pages 89–98, 2003.

- [19] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56:9–33, 2004.
- [20] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M.L. Loh, J. R. Downing, M. A. Caliguri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.
- [21] V. M. Govindu. A tensor decomposition for geometric grouping and segmentation. In *IEEE CVPR*, 2005.
- [22] W. H. Greub. *Multilinear Algebra*. Springer, 1967.
- [23] J. A. Hartigan. *Clustering Algorithms*. Wiley, 1975.
- [24] M. Hein and O. Bosquet. Hilbertian metrics and positive definite kernels on probability measures. In *AISTATS*, 2005.
- [25] Y. Kluger, R. Basri, and J. T. Chang. Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Research*, 13:703–716, 2003.
- [26] T. G. Kolda and J. Sun. Scalable Tensor Decompositions for Multi-aspect Data Mining. In *ICDM*, 2008.
- [27] A. Kumar, Y. Sabharwal, and S. Sen. A simple linear time $(1 + \epsilon)$ -approximation algorithms for k-means clustering in any dimensions. In *IEEE Symp. on Foundations of Comp. Sci.*, 2004.
- [28] S. P. Lloyd. Least squares quantization in PCM. *IEEE Tran. on Inf. Theory*, 28(2):129–136, 1982.
- [29] B. Long, X. Wu, and Z. Zhang. Unsupervised learning on k-partite graphs. In *SIGKDD*, 2006.
- [30] R. Nock, P. Luosto, and J. Kivinen. Mixed bregman clustering with approximation guarantees. In *Euro. Conf. on Mach. Learning (ECML)*, LNAI 5212, 2008.
- [31] K. Puolamäki, S. Hanhijärvi, and G. C. Garriga. An approximation ratio for biclustering. *Inf. Process. Lett.*, 108(2):45–49, 2008.
- [32] I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of American Mathematical Society*, 44(3):522–536, 1938.
- [33] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2001.
- [34] A. Shashua, R. Zass, and T. Hazan. Multi-way Clustering Using Super-Symmetric Non-negative Tensor Factorization. *LNCS*, 3954:595–608, 2006.
- [35] S. Sra, S. Jegelka, and A. Banerjee. Approximation algorithms for bregman clustering co-clustering and tensor clustering. Technical Report 177, MPI for Biological Cybernetics, Oct. 2008.
- [36] H. Zha, C. Ding, T. Li, and S. Zhu. Workshop on Data Mining using Matrices and Tensors. *KDD*, 2008.

